Udacity Data Science Challenge

Sundeep Pattem

This report presents work on the 'Udacity Data Science Challenge' based on student activity data provided for the CS-255 course. A summary of results is followed by a description of the dataset, details of how the results were arrived at, and an appendix with additional results.

I. Summary of results

A summary of insights and recommendations is as follows:

- **Session definition**: Compute cumulative activity (node visits + submissions) for every time period T since enrolment. A continuous sequence of hours with non-zero activity is considered a session. When two non-zero sequences are separated by a single zero, combine and consider as a single session. The data suggests that T = 1 hour would be a balanced choice for this course.
- **Session stereotypes**: Sessions of students achieving different mastery levels can be characterized by the distribution of session durations and the submission frequency within a session. High levels of mastery show an even distribution of short and long sessions and low proportion of sessions with no submissions and high proportion of moderate and low submissions. Lower or no mastery levels show a high proportion of short sessions and sessions with no submissions
- **Tracking student progress and predicting outcomes**: Individual student outcomes correspond well with cumulative daily (a) ratio of submissions to node visits and (b) number of submissions evaluated as correct/True. Results show that a progress metric that combines the two is strongly indicative of higher, lower and no mastery levels achieved by students in as few as 5 days after enrollment.
- **Content bottlenecks**: 50% of students who achieve level 1 stop progressing at lesson 'l-49464373' and a few of its exercises. The instructors can consider if there is a possibility for restructuring this lesson and its exercises to help more students to make further progress in the course.

Further work of immediate interest is:

- an analysis of the amount of time spent on particular lessons, exercises, and morsels and the number of correct/incorrect (True/False) submission attempts by successful and unsuccessful students and

- to use features shown to be predictive of learning outcomes (session durations, submissions per hour of session, ratio of submissions to node visits, rate of correct/True submissions) to build a more discriminative classifier.

II. Dataset description

The challenge is to engineering features for classification of student ”learning sessions,” which intuitively are contiguous units of student activity in the Udacity classroom with only a few short breaks allowed. Given a sampling of activity for a population of students, these are some questions of interest:

- How do we define learning sessions (i.e. given a time series of student activity data, how can we mark the end of one session and the start of another), and what features can we see distinguishing one kind of session from another?
- Do students tend to have sequences of similar session stereotypes?
- Can we find any changes in student behavior that may predict when a student is falling off pace? Do any features strongly predict a particular learning outcome?

A. Data Sample

The sample data set includes three files:

1) **accounts.csv:** A list of accounts who enrolled in our CS255 HTML5 Development course in June, 2013. The file contains the following fields:
   - account_key: A unique alpha-numeric string representing the Account (key value has been replaced with a pseudonym to protect the innocent)
   - enrollment_time: A UTC-zoned ISO8601 date-time string representing when the Account registered for CS255
   - learning_outcome: The level of ”Mastery” certificate the Account has obtained to date for CS255. The maximum level is 4. An empty value means no certificate has been earned.
2) **nodevisits.csv:** All the recorded NodeVisits for the accounts included in accounts.csv. A NodeVisit represents when a content node (video or quiz) in our course tree structure is presented to the student. One may think of NodeVisits as roughly equivalent to a page view. The file contains the following fields:
   - activity_id: A unique identifier for the Activity
   - account_key: The Account to which the Activity belongs

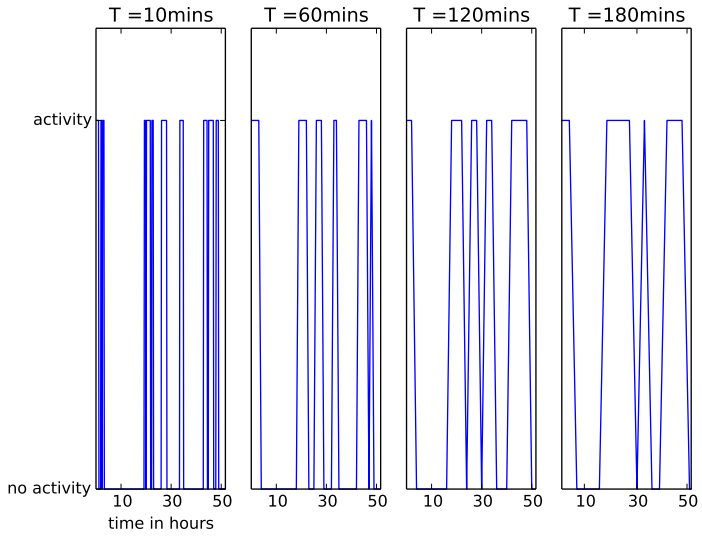


Fig. 1. The same sequence of activity is sought to be sessionized with different values of T. Low values lead to closely spaced spikes, while large values subsume boundaries. Sequences of continuous activity with T= 60 mins seen to be reasonable choice. More samples in Appendix.

- time: A UTC-zoned ISO8601 date-time string representing when the activity occurred. Note that sequencing from this date is possible but is subject to clock sway and network latency.
- content_path: A hierarchical path representation for the location of the content node visited. This will match the portion of the url seen in the browser's address when inside the classroom.

3) submissions.csv: All the recorded quiz submissions for the accounts included in accounts.csv. The file contains the following fields:

- activity_id: A unique identifier for the Activity
- account_key: The Account to which the Activity belongs
- time: A UTC-zoned ISO8601 date-time string representing when the activity occurred. Note that sequencing from this date is possible but is subject to clock sway and network latency.
- content_path: A hierarchical path representation for the location of the content node visited. This will match the portion of the url seen in the browser's address when inside the classroom.
- evaluation: A nullable boolean string ('True', 'False', ' '), where 'True' means the submission was correct, 'False' means the submission was incorrect, and ' ' means the submission was recorded by clicking the "Test" button on the UI and thus no evaluation was performed.

III. Session definition

Intuitively, for an individual user, activity separated by a few hours can be considered to be in different sessions and when separated by a few minutes can be taken to belong to the same session. The question then is about where the line should be drawn, and to find a useful characterization for what is "few" in this situation.

Consider a time interval of duration T. For every user, starting with the day of enrollment, the total duration of activity is divided into periods of duration T and cumulative activity (node visits + submissions) for each period is computed. A continuous sequence of periods with non-zero activity is considered a session. An illustrative pattern is shown in Figure 1, where sessions appear as spikes/trapeziums. When T is too small compared to actual session durations, this can be expected to result in a bunching together of activity spikes. When T is too large, (a) a similar pattern is observed as the beginnings and ends of actual sessions get arbitrarily assigned to adjacent periods, and (b) activity separated by long periods gets included in the same period.

For a number of student sessions (see Appendix), it is observed that setting T = 1 hour is a reasonable choice, and leads to a relatively smaller number of adjacent non-zero sequences. When two non-zero sequences are separated by a single zero, they can be combined and considered as a single session. Illustration with T = 1 hour for samples of students achieving different mastery levels in first 100 hours since enrollment. See Appendix for more samples of student sessions across different mastery levels achieved.

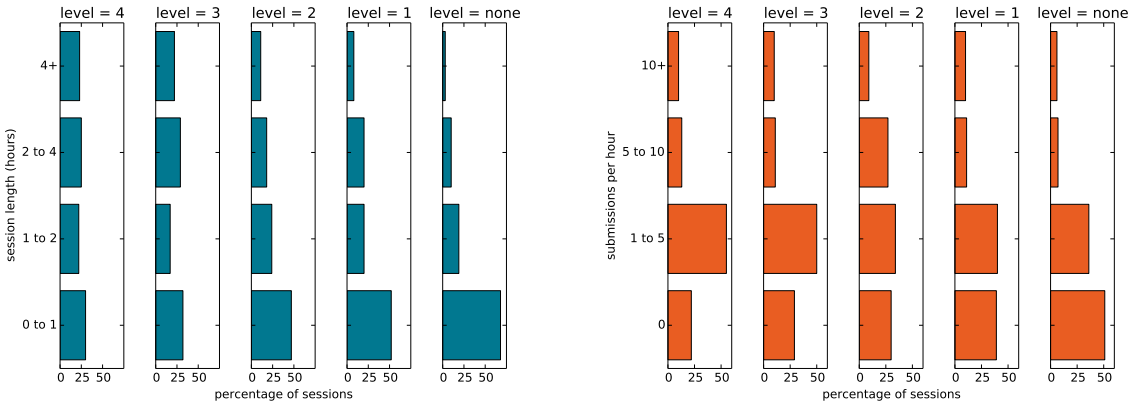
IV. Session stereotypes

We have 5 classes of students/outcomes, with the achieved mastery levels of 4 (10 students), 2 (7 students), 3 (15 students), 4 (108 students) and none (5800). We expect that learning corresponds to time spent on the material, the amount of total activity (node visits + submissions), and that the level of the assigned learning outcome will depend on the number of quiz questions attempted/submitted. Given our characterization of the session above, we can now investigate the distribution of session durations and frequency of submission activity.

In Figure 2, we see that students achieving:

(i) high levels of mastery have (a) an even distribution of short and long sessions and (b) low proportion of sessions with no submissions and high proportion of moderate and low submissions.

(ii) lower or no mastery level have a high proportion of (a) short sessions and (b) sessions with no submissions



(a) Distribution of session lengths. (b) Submissions made per hour in session.

Fig. 2. Students achieving high levels of mastery have (a) an even distribution of short and long sessions and (b) low proportion of sessions with no submissions and high proportion of moderate and low submissions. Those achieving lower or no mastery level have a high proportion of (a) short sessions and (b) sessions with no submissions

These results show that the distribution of session lengths and the distribution of submissions per hour in session are promising features for predicting learning outcomes. It is also possible to classify individual sessions based on the ratio of node visits and submissions, and the success (correct/True evaluations) rate of the submissions. We look at these next.

V. Tracking student progress and predicting outcomes

To get an overview of behavior of students over the duration of the course, we look at the cumulative daily activity. Since both node visits and submissions drop sharply around day 30 (see Appendix), we surmise that CS-255 was of a 4-week or 1 month duration, and consider activity up to 40 days.

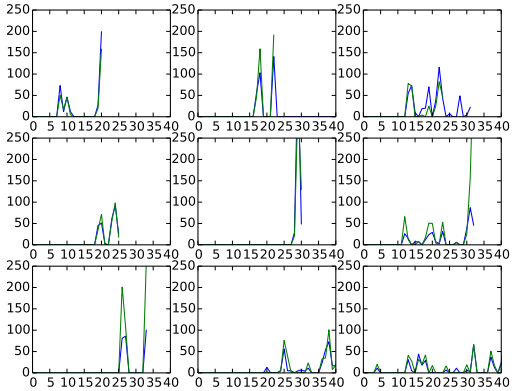
Figure 3 shows that the

(i) daily ratio of submissions to node visits corresponds well with mastery level. Students with consistently high ratios achieve high mastery levels.

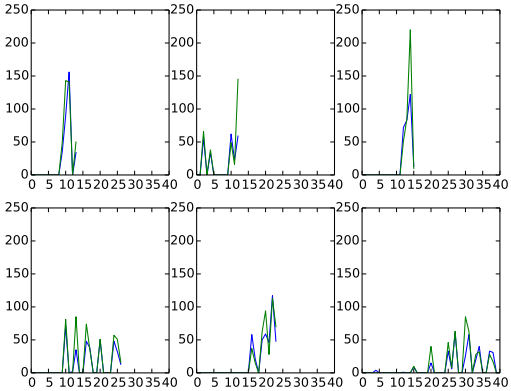
(ii) most successful students either have consistently moderate activity or peaks of very high activity

(iii) less successful students either have moderate activity for limited periods or one peak and very little activity otherwise.

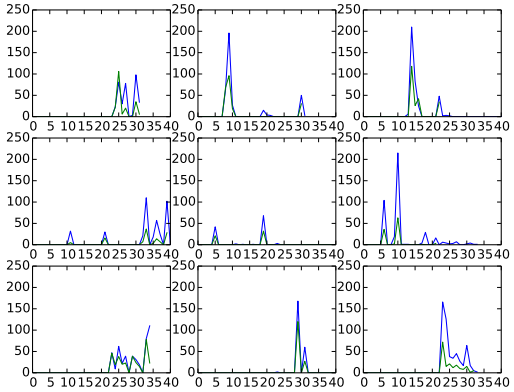
To predict individual outcomes, we further investigate the daily ratio of submissions to node visits. In addition to having a higher number of submissions, it can also be expected that students making good



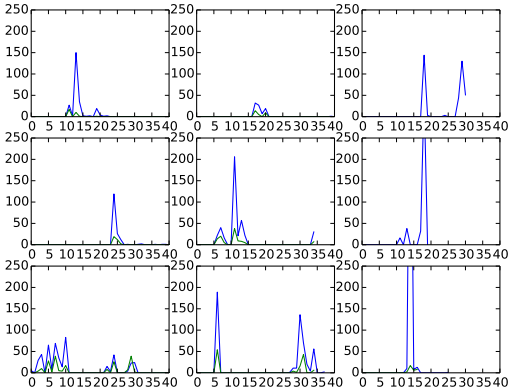
(a) mastery level = 4



(b) level = 3



(c) level = 1



(d) no level achieved

Fig. 3. Cumulative daily activity for 40 days (course duration). X-axis is day since start of course, Y-axis is the cumulative activity for the day with node visits in blue and submissions in green. Each box corresponds to an individual student. It is seen that the ratio of submissions to node corresponds well with mastery level. Also, higher mastery levels show either consistent moderate activity or very high peaks, while other show less consistency and isolated peaks of node visits.

progress in learning will have a higher number of submissions evaluated as correct (True). To capture the above, we define a progress score:

$$P(d_n) = \sum_{d=d_0}^{d_n} CS(d) \cdot \frac{S(d)}{NV(d)} \tag{1}$$

where d_0 is the day the course begins, d_n is nth day since the course began, CS(d) is the number of correct submissions on day d, $S(d)$ is total submissions and $NV(d)$ is the number of node visits.

Figure 4 shows the progress score (in log scale) for individual students. The x–axis is time, normalized

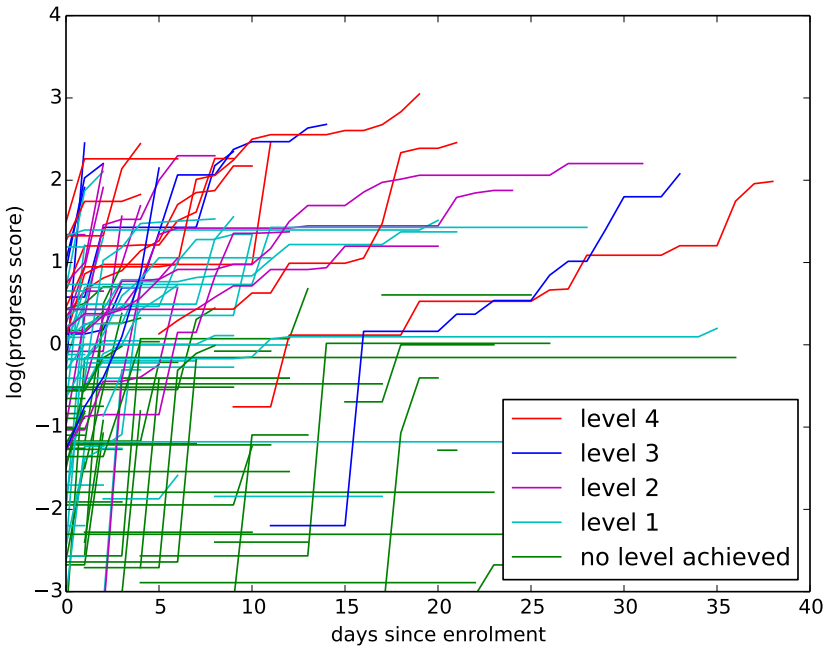


Fig. 4. Evolution of progress score $P(d)$ for individual students as a function of days since enrolment. Typical profiles seen for students achieving higher, lower and no mastery levels. The score can provide a fair prediction of learning outcomes after as little as 5 days past enrollment/start of significant activity.

to be the number days since enrollment for each student. This can provide even more clarity if adjusted for date of first significant activity.

We observe that most students who achieve:

- some level of mastery can be identified within the first 5 days.
- mastery levels of 3 and 4 have two typical profiles: either a sharp and significant rise in the first 5 days or a steady rate of moderate progress over a longer duration.
- mastery levels of 1 and 2 show early moderate progress which then levels off.
- no mastery level show little progress at any time (note log scale).

VI. Bottlenecks in course content

The course material visited by at least one student consists of 12 lessons, 81 exercises and 278 morsels. We investigate if there are particular cliffs/bottlenecks in the course material where students usually struggle/drop off. To obtain the sequence of lessons in the course, we extract the order in which

lesson_id	avg_order_visited	num_exercises	num_morsels	num_submit_morsels
l-74901984	1.00	0	2	0
l-52473341	2.01	7	28	6
l-66584787	2.97	5	19	4
l-undefined	3.47	6	14	0
l-49464373	4.00	10	42	10
l-52850040	4.89	6	26	6
l-52265917	5.52	17	59	17
l-185058926	5.66	1	1	1
l-52850041	5.77	10	36	10
l-52850042	6.05	5	21	5
l-52742191	6.79	9	31	9
l-185534958	10.00	12	12	12

TABLE I

AVERAGE ORDER OF LESSONS VISITED FROM TOP TO BOTTOM, AND THE NUMBER OF EXERCISES, MORSELS, AND SUBMITTABLE MORSELS IN EACH LESSON.

content_path	num_students_final_activity
l-49464373/e-73862317/m-93053898	9
l-49464373/e-73862318/m-101628096	8
l-49464373/e-73862319/m-93053899	6
l-49464373/e-73862320/m-101628097	5
l-49464373/e-73862321/m-101628098	6
l-49464373/e-73862322/m-101628099	4
l-49464373/e-73862333/m-101628100	5

TABLE II

EXERCISES AND MORSELS IN LESSON 'L-49464373' WHERE MOST STUDENTS GET STUCK AND FALL BEHIND

each student visits the lessons and average the rank for each lesson over all students who visit it. Table I shows the sequence of the lessons, and the number of exercises, morsels, and submittable morsels in each lesson.

Figure 5 shows the proportion of last lessons visited by students achieving different mastery levels. For now, we ignore the cases where students refer back to earlier material after the level is assigned. As expected, a large number of students fall off in the very first two lessons and do not achieve a mastery level, while a major proportion of those achieving higher levels end in the last few lessons.

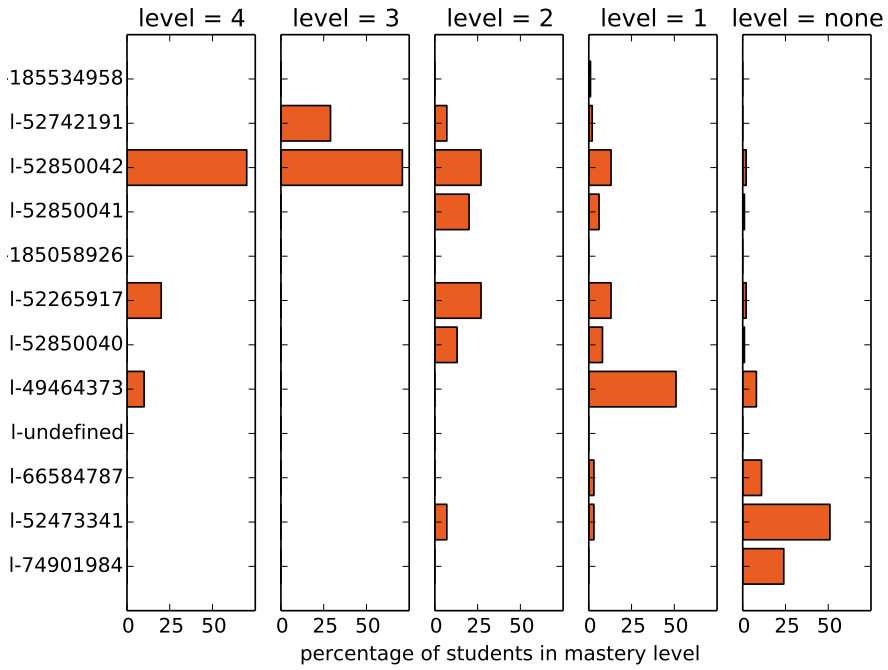


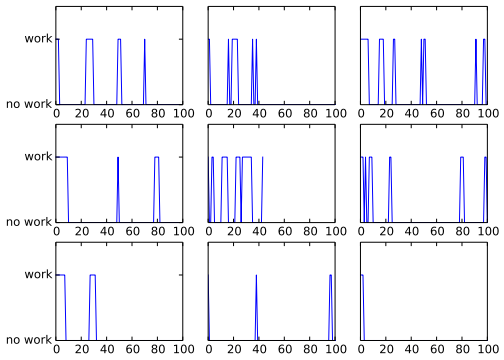
Fig. 5. Proportion of last lessons visited by students achieving different mastery levels, Interestingly, 50% of students who achieve mastery level 1 stop progressing at lesson 'l-49464373'.

The standout figure is that 50% of students who achieve level 1 stop progressing at lesson 'l-49464373'. From Table I, we see that this is the first 'heavy' lesson with 10 exercises, 42 morsels and 10 submittable morsels. The exercises and morsels that most of these students end/fall behind at are listed in Table II.

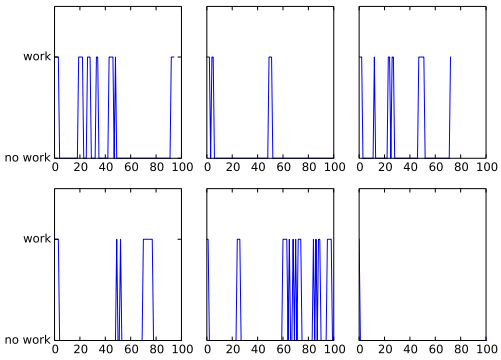
The instructors can consider if there is a possibility for restructuring this lesson and its exercises to help more students progress and work on material deeper into the course. Analysis of the amount of time spent on particular lessons/exercises/morsels and the number of correct/incorrect (True/False) submission attempts by successful and unsuccessful students can provide further insights.

VII. Conclusion

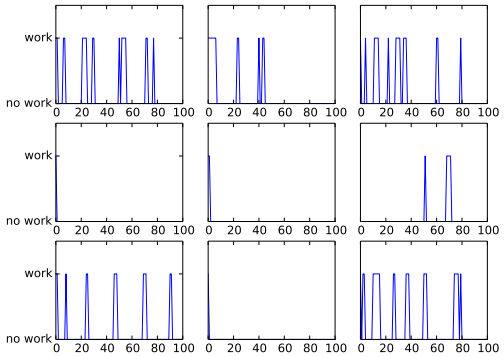
A workable definition of a session for this course was obtained from the data. Using this notion of a session, the distributions of session durations and rate of activity in the session were shown to correspond well with student categories by mastery levels achieved. An interesting observation from the daily cumulative activity for individual students over the duration of the course is that the ratio of submissions to node visits is strongly indicative of learning outcomes. This ratio was combined with



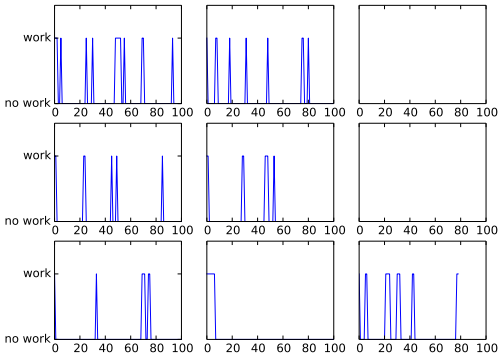
(a) mastery level = 4, T = 60 mins



(b) level = 3, T = 60 mins



(c) level = 1, T = 60 mins



(d) no level achieved, T = 60 mins

Fig. 6. Session determination for samples of students achieving different mastery levels in first 100 hours since enrollment. Each box corresponds to an individual student.

the rate of correct/True submissions to design a progress score metric which is able to predict learning outcomes for students in as few as 5 days following enrollment/significant activity. Analysis of the progress through the course content revealed a bottleneck at which a significant number of students were falling behind/giving up. The course designers can consider options for helping students overcome this stumbling block.

While the score developed is fairly promising, and samples for higher mastery level outcomes are very few, it should be possible to build a classifier with all the features identified as input for better discrimination. Closer analysis of the time spent by both successful and unsuccessful students on portions of the course might provide more insights for improving the analysis and the course.

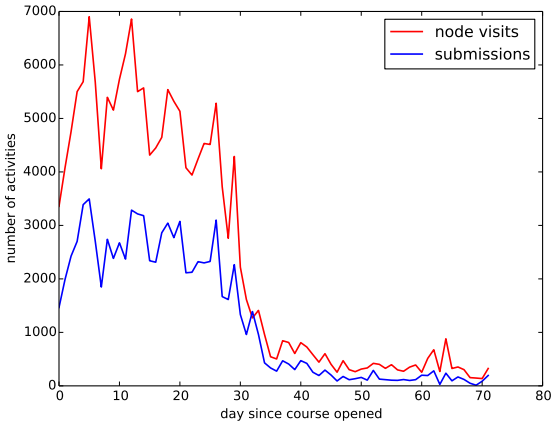


Fig. 7. Cumulative daily activity for CS-255.

VIII. Appendix

A. *Session durations*

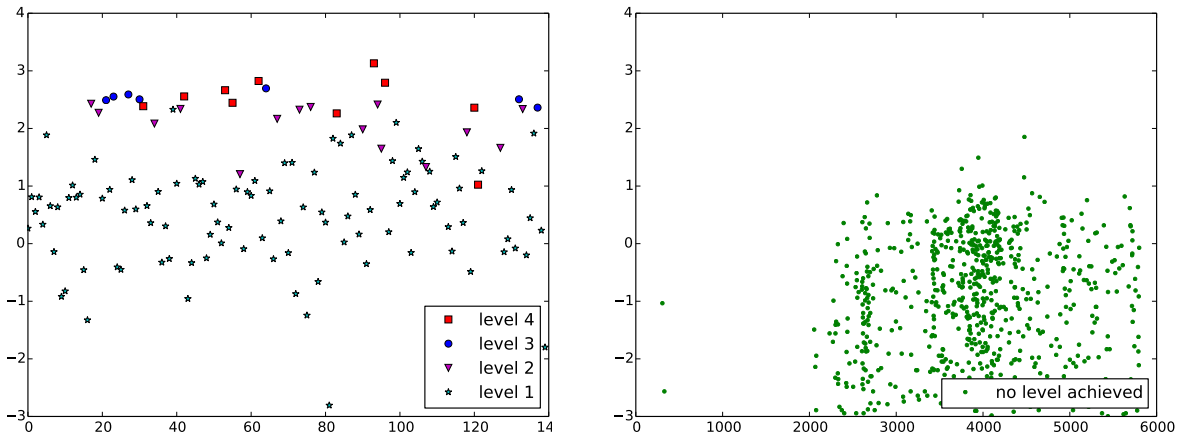
Figure 6 shows that the choice of T = 1 hr for determining sessions is a reasonable choice across students with different profiles and outcomes

B. *Daily activity over course duration*

Figure 7 shows the cumulative daily activity for the 72 days for which data is available. Since both node visits and submissions drop sharply around day 30 (see Appendix), we surmise that CS-255 was of a 4-week or 1 month duration, and consider activity up to 40 days in most analysis.

C. *Progress score for learning outcome discrimination*

Figure 8 shows that the progress score has potential for discriminating between students with higher and lower or no mastery level achieved. For more granular separation, a classifier that takes the features shown to have predictive power as input is of immediate interest.



(a) (b)

Fig. 8. Progress score values at end of course. Students with mastery level > 1 can be well separated from those with lower and no mastery levels.